GHAJAR
EXHIBIT 30

Page 1

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

SAN FRANCISCO DIVISION

```
______________________________
RICHARD KADREY, et al.,       )
                              )
          Individual and      )
          Representative      )
          Plaintiffs,         )
                              )
v.                            )   Case No. 3:23-cv-03417-VC
                              )
META PLATFORMS, INC.,         )
                              )
          Defendant.          )
______________________________)
```

** HIGHLY CONFIDENTIAL - ATTORNEYS' EYES ONLY **

Videotaped Deposition of SEAN BELL

San Francisco, California

Wednesday, December 11, 2024

Reported Stenographically by

Michael P. Hensley, RDR, CSR No. 14114

DIGITAL EVIDENCE GROUP
1730 M. Street, NW, Suite 812
Washington, D.C. 20036
(202) 232-0646

Page 12

Q. All right. And what's your position there?

A. I'm a research scientist manager.

Q. All right. And when did you start at Meta Platforms?

A. In January 2019.

Q. And when you started at Meta Platforms in 2019, did you hold the same position that you hold today?

A. Yes. I -- my job offer may have said something different. I don't remember the exact title, but I've -- I've been a manager since I started.

Q. Okay.

A. It may have said "research scientist," for example.

Q. Have you been promoted since you began working at Meta?

A. Yes. I was promoted from Level 6 to Level 7.

Q. And what does that mean to be promoted from Level 6 to Level 7?

Page 24

A. Yes. That's one of the things that I do.

Q. Okay. And you said that you work on -- you also work on collecting data for other foundation models.

Did I hear that correctly?

A. That's right.

Q. What other foundation models do you collect data for?

A. So I'm currently collecting data for Llama 4 for the -- so that's the first one. The second category is the image generation models. We called it "Emu," E-m-u. And then the third category is the movie generation models called "Movie Gen."

And so sometimes the data that my team collects, even though it was intended for -- primarily for one of those three categories of models, it might be useful to other teams; and so occasionally we'll establish collaborations where we might additionally share data with other teams.

Q. Okay. And in your role working on the Llama models, do you provide any strategic guidance on collecting or using data?

Page 25

A. Yes.

ATTORNEY HARTNETT: Object to the form.

THE WITNESS: Yes.

///

BY ATTORNEY STOLER:

Q. Okay. What sort of guidance do you provide as to the use of data for training Llama?

A. So ultimately what my team is accountable for is the final dataset that goes into Llama. And what -- what matters the most is, you know, what is the benchmark performance as a result of training on this data?

And so our goal is to maximize the benchmark performance as a result of training on the data, and so there's a huge number of different steps prior to that point, prior to the actual data you put into Llama.

And so I might be consulting on any part of the strategy that leads up to that point. For example, how do you process data? How do you curate data? How do you measure data quality? How can you decide that one dataset is better than another

Page 26

dataset? You know, what are the -- what is even the scientific method to know how to attribute, you know, this concept of, quote/unquote, "quality"? Which means, really, the benchmark performance as a result of training on the data. You know, it may also be what -- you know, thinking about what kinds of sources we need to be collecting.

And so generally -- either I might be writing some of the strategies, or people -- or leadership on my team may be writing those strategies and then I'm reviewing it.

BY ATTORNEY STOLER:

Q. And prior to coming to Meta, did you ever work on large language models?

A. No.

Q. So fair to say you learned about data strategy for LLMs while at Meta?

A. Yes.

Q. Okay. And who did you learn that from?

A. For LLMs specifically, I read a lot of academic papers. And also I -- you know, as a manager, I took on a number of senior individual

Page 41

BY ATTORNEY STOLER:

Q. Let me rephrase.

A. Yeah.

Q. Does the quality of the pretraining data affect the performance of an LLM?

ATTORNEY HARTNETT: Object to the form.

THE WITNESS: So what I was trying to get at earlier is that we've basically defined our notion of quality to be the performance of the LLM, and so the benchmarks are a way to estimate the -- you know, estimate the performance.

You know, this was actually a major challenge and question that we had to think about when we started the team, which is how do you think about quality and is it possible to know what it is. And so we had some heuristics but we didn't have good science connecting it.

And so a lot of the work that this team is doing is trying take something that's a little bit subjective and turn it into something as scientific as possible, meaning that, you know, there's no sort of judgment around it. It's a connection between a

Page 42

particular experiment that you do and then a very measurable set of numbers.

(Clarification inquiry by The Court Reporter.)

///

BY ATTORNEY STOLER:

Q. What does the term -- in the context of pretraining LLMs, what does the term "data mix" refer to?

A. So "data mix" refers to the final overall combination of data that the LLM is actually trained on. And so there's really two kinds of activities when creating -- when collect -- when creating a dataset.

One of them will be on an individual dataset basis, which types or sources of data do we want to use, and which subset of the data -- or how do I filter it or process it? And so that may give us, say, order of approximately 100 different individual sources.

The data mix is an additional step where I now need to choose the relative weighting or

Page 43

proportion of those individual datasets. And so if the weight is particularly high, I might repeat certain datasets more often. If the weight is less than 1, for example, then I might be taking a proportional subsets; maybe I'll sample only 80 percent of some data source, for example.

And so the data mix is the work of deciding all of those weights and proportions and maybe whether or not to include a dataset at all.

Q. Do you oversee the determination of the data mix for Llama 4's pretraining data?

A. Yes.

Q. Help me understand.

You testified that you are not aware of what data goes into -- sorry, let me -- I'll start again.

Help me understand. Did you previously mention that you don't know what datasets were used to train Llama 3 but are not being used to train Llama 4?

ATTORNEY HARTNETT: Objection to form. Asked and answered.

Page 49

Llama 4 pretraining dataset, would that render your picture of the training data incomplete?

ATTORNEY HARTNETT: Objection to the form.

THE WITNESS: I'm not sure what you mean by that.

BY ATTORNEY STOLER:

Q. Yeah. If you don't know what datasets went into pretraining Llama 4, would you be able to evaluate whether the dataset you created was optimal?

ATTORNEY HARTNETT: Objection to the form.

THE WITNESS: So we -- so a lot of what we're focused on -- I mean, I'm -- honestly, I'm not sure that it's needed.

So I mean, we -- what we do is we -- we've spent a lot of time figuring out what is the science of what makes for good data in terms of the benchmark performance. And so we have done a lot of innovations in this space and come up with new types of metrics and ways to understand the data.

And some of what we've learned in the past is that human intuition on data is -- is wrong. And

Page 50

so in 2023, when I was not there, but I -- you know, other researchers --

I'm sorry. My microphone dropped. Let me put it back, sorry.

In 2023, I was told about research projects where people attempted to look at the data and make human judgements just on looking at the data, did they think it would be useful to LLM training or not.

And in a different workstream, they said let's take a statistical approach where we only look at the benchmark performance, and what they found was that the human intuition actually led you to a worse-performing LLM in the end and that it was counterproductive to attempt to, on a per-document or per-dataset basis, make human judgments about that.

And so when I say that I've "shifted," it was based on scientific learning; that the best way to do data mix optimization is to, as much as possible, distill everything into a statistical view.

Page 51

Again, when I say "statistical view," I mean, like, the distribution of tokens, the distribution of benchmark scores, and, you know, look at that view as much as possible, make all your decisions in that lens. And then separately have a central private -- you know, a privacy infrastructure that guarantees that we're following all of the -- all of our mitigation decisions and that -- you know, one of the principles I've -- I've pushed the team building that system was this idea that you couldn't train on bad data even if you tried. And when I say "bad data," I mean data that disagrees with any of our privacy decisions.

BY ATTORNEY STOLER:

Q. You mention human -- let me back up.

You just said that there was a switch in the approach that you took to evaluating data.

Did I get that correct?

A. I wouldn't call it a switch. And so, again, I -- I wasn't there; so I've only heard, you know, small pieces of what happened.

Q. Hmm.

Page 52

A. But my understanding, hearing from others, is that there is a team in Europe that questioned whether or not the team in the U.S. was doing the right thing in terms of pretraining data. And so they started a separate project to do human annotation of data to try and use human judgments to label each dataset -- is this a good quality dataset or not? -- and, you know, build a whole process around this.

You know, basically what happened was the team in Europe looked at the pretraining data directly, and they -- they thought, "This is very low-quality data. You guys don't know what you're doing. How -- why are we training on such bad data from this human intuition?"

And so they started a project to do annotation of this data to then label data as "good quality" or "bad quality." And then it turns out that that whole approach of annotation did not lead to more useful filters, and so we're not using that approach.

So it wasn't so much that we switched

Page 53

approaches. It was that there were a lot of different approaches being tried at the same time, and one of them was this annotation process to challenge the results of the other team.

You know, and the last thing I'll add is that I wasn't here for any of this, and so this is me, you know, recalling what -- how other people described to me how it went.

Q. And what you just described, was that happening in connection with the pretraining of Llama 3?

A. That is my understanding.

Q. Do you know if it was taking place before Llama 3 was pretrained or after Llama 3 was pretrained?

ATTORNEY HARTNETT: Objection to the form.

THE WITNESS: So this took place in -- this took place during the process of deciding on the pretraining data for Llama 3, is my understanding.

BY ATTORNEY STOLER:

Q. Okay.

Page 64

text data within LibGen?

A. No. I don't know.

Q. All right. Have you ever heard of a dataset referred to as "Anna's Archive"?

A. Yes. I've heard of it.

Q. Is Anna's Archive a books dataset as well?

ATTORNEY HARTNETT: Object to the form.

THE WITNESS: So my understanding is that Anna's Archive is a superset of LibGen and contains a number of things. And so I -- my understanding is that it contains books, but it -- it also has a lot of other things in there.

BY ATTORNEY STOLER:

Q. Do you know if Meta ever used the Books3 dataset to train any of its Llama models?

A. Well, I've seen it mentioned in the Llama paper; so for Llama 1. I don't know what data was trained for Llama 2.

So I -- there was a list of code names for Llama 3, and so I -- I don't actually know which of the code names corresponds to Llama 3. So it may be in there, but I don't know for certain.

Page 97

anyone in your team regarding the importance of training on the Anna's Archive dataset?

ATTORNEY HARTNETT: Objection to form.

THE WITNESS: That's pretty vague. What do you mean by that?

BY ATTORNEY STOLER:

Q. Has anyone on your team ever indicated to you that it's important that Llama 4 train on the Anna's Archive dataset?

A. I mean, that would've been connected with us expanding from LibGen to Anna's Archive.

And so the general understanding was that we need more reasoning data. "Reasoning" broadly meaning STEM and mathematics and that one of the best sources for reasoning data was in LibGen. And so for that reason, you know -- then there was a question of, you know, will we use all of LibGen or not and do we need to look at other sources or related things similar to LibGen which was around the conversation of saying, you know, let's also think about Anna's Archive.

Q. My question is just whether you remember

Page 131

mean?

A. Like, downloading it from publicly available sources --

Q. Okay.

A. -- in the context of this bullet here.

Q. And can you think of an example of a publicly available dataset of books/PDFs that you downloaded?

A. Sure. I mean, there is a whole bunch of PDFs on Spidermate and web crawling. And so, you know, we made sure that when we're working with that team that we ask them specifically to make sure to prioritize downloading PDFs.

Q. Okay. And how is -- so I see that sourcing books and PDFs is separate from licensing.

The -- is there -- do you know if there was any attempt to license any of these books or PDF datasets?

A. So --

ATTORNEY HARTNETT: Object to the form.

THE WITNESS: So at this time, we -- we had -- if I remember correctly, we had not really

Page 132

gotten into trying to license anything at scale. And, you know, first we needed budget and sort of leadership alignment around the very idea that -- should we be approaching publishers or not.

Since then, we started approaching some publishers, and what we found was that the volume of content that they have is orders of magnitude smaller than what we needed for pretraining; and so that changed the focus from "Okay. Should we think about licensing as a part of strategy of our pretraining?"

And so there was a shift in thinking from the time when we wrote this to now, which is that, "Hey, licensing. We should think of it as a -- something for posttraining."

Like, it just doesn't exist in the scale that you need for pretraining to even be meaningful. Like, whether or not you add -- like, say hypothetically you licensed every book from some publisher and tried to put it into pretraining. It would be a statistical noise that would be not even measurable of whether or not you did it --

Page 133

pretraining scale, in terms of the volume of the content.

And so -- you know, so since then, we've been focusing on licensing for posttraining. And so because I'm focused mostly on pretraining, I have not been involved -- that involved in licensing.

BY ATTORNEY STOLER:

Q. When you say that it would be -- sorry. When you say that it would amount to "statistical noise" to include every book in the pretraining data, what do you mean by that?

A. So what I mean is that the minimum size right now we're able to measure, approximately speaking, is about 100 billion tokens. So the final data mix is, you know, say, 30 trillion in some of the models and 60 trillion we're working towards for the Llama 4 flagship model.

And the benchmarks themselves already are a little bit noisy. So there's already some plus or minus on certainty. And so if you change the data, you have to know whether or not the change in data resulted in a benchmark movement larger than the

Page 136

A. Well, I mean, the scale that we're able to get in terms of PDFs -- and there's a lot of books in those PDFs -- is something that we can measure, you know, in terms of what you can download in terms of publicly available sources.

So I'm talking about specifically a licensing strategy where there is a bit of a song and dance with, you know -- so say we go to someone who has content to license and, you know, there's a -- you know, haggling process on the price. You know, it takes a bunch of their time; it takes our time. They might send us a sample of the data. They don't necessarily want to the give us all data because then they think, "Well, I've just given it to you."

So now -- so the problem is compounded even more that they might give us a tiny sample of data, which is very hard to figure out what it -- so the problem of figuring out the value is made 100 times harder because they give us a tiny sample.

And as I mentioned earlier, what we know is that human intuitions studying an individual

Page 304

CERTIFICATE OF SHORTHAND REPORTER

I, Michael P. Hensley, Registered Diplomate Reporter for the State of California, CSR No. 14114, the officer before whom the foregoing deposition was taken, do hereby certify that the foregoing transcript is a true and correct record of the testimony given; that said testimony was taken by me stenographically and thereafter reduced to typewriting under my direction; that reading and signing was not requested; and that I am neither counsel for, related to, nor employed by any of the parties to this case and have no interest, financial or otherwise, in its outcome.

Michael P. Hensley, CSR, RDR